Explainable Machine Learning: Understanding the Limits & Pushing the Boundaries

Hima Lakkaraju

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- Towards Robust & Reliable Explanations
- The Road Ahead

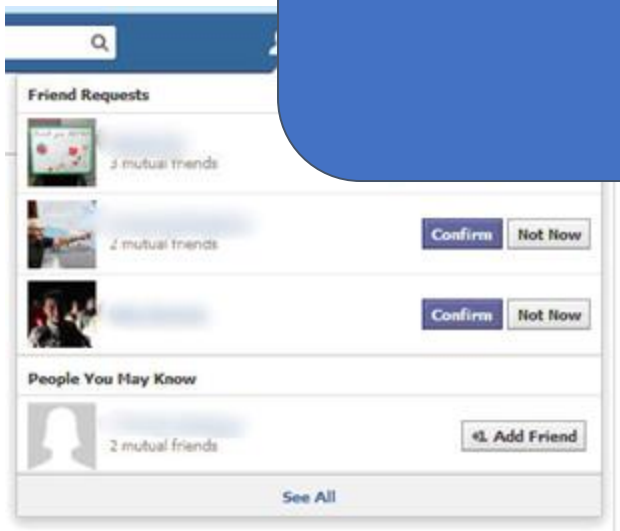
Tutorial Outline

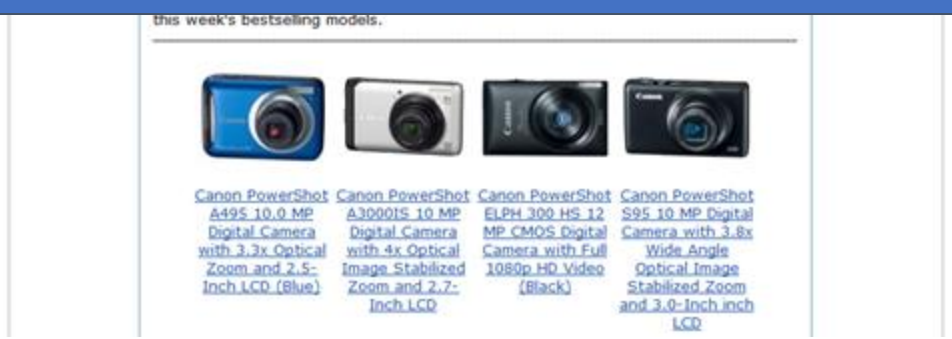
- **Motivation**
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- Towards Robust & Reliable Explanations
- The Road Ahead

[Weller 2017]

Motivation

Machine Learning is EVERYWHERE!!





Motivation

Model understanding is absolutely critical in several domains -- particularly those involving *high stakes decisions*!

Motivation: Why Model Understanding?

Input

This model is relying on incorrect features to make [illegible]iction!! Let [illegible] model

Model understanding facilitates debugging.

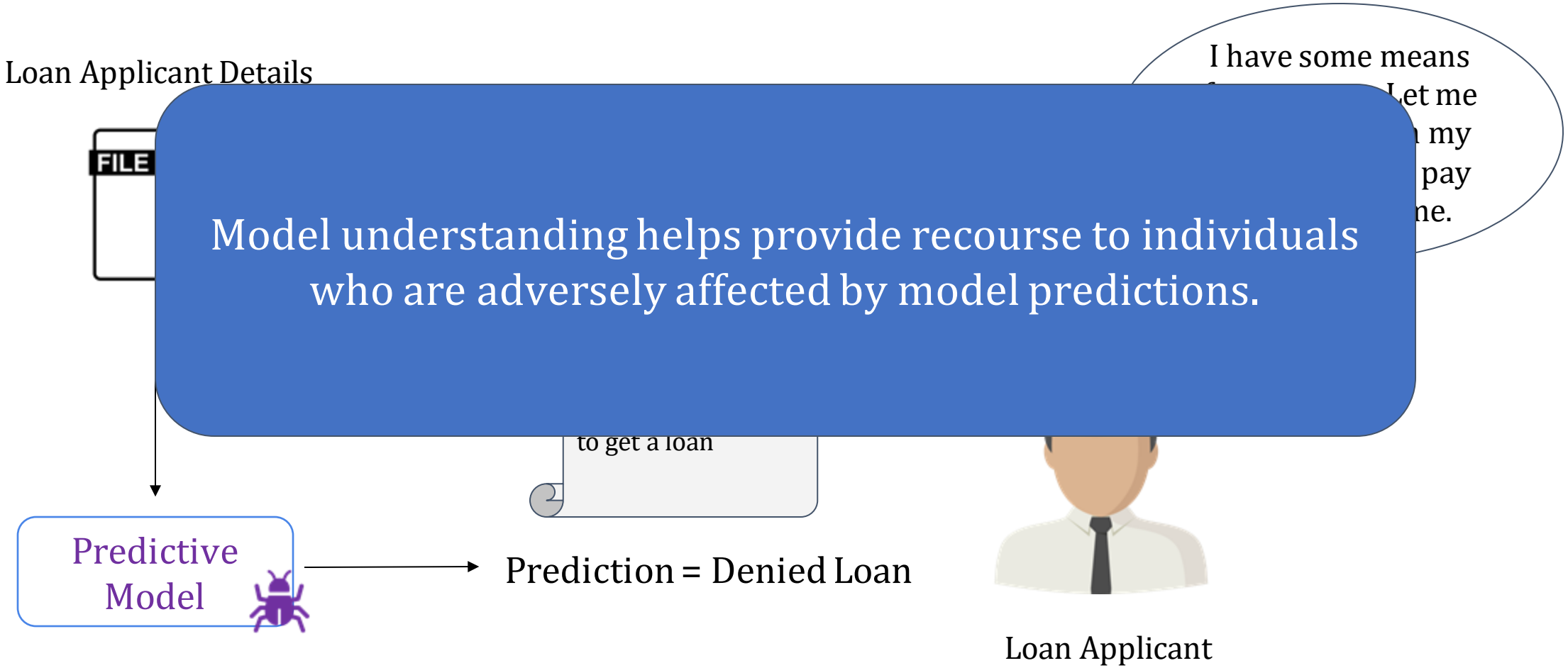
Predictive Model

Prediction = Siberian Husky

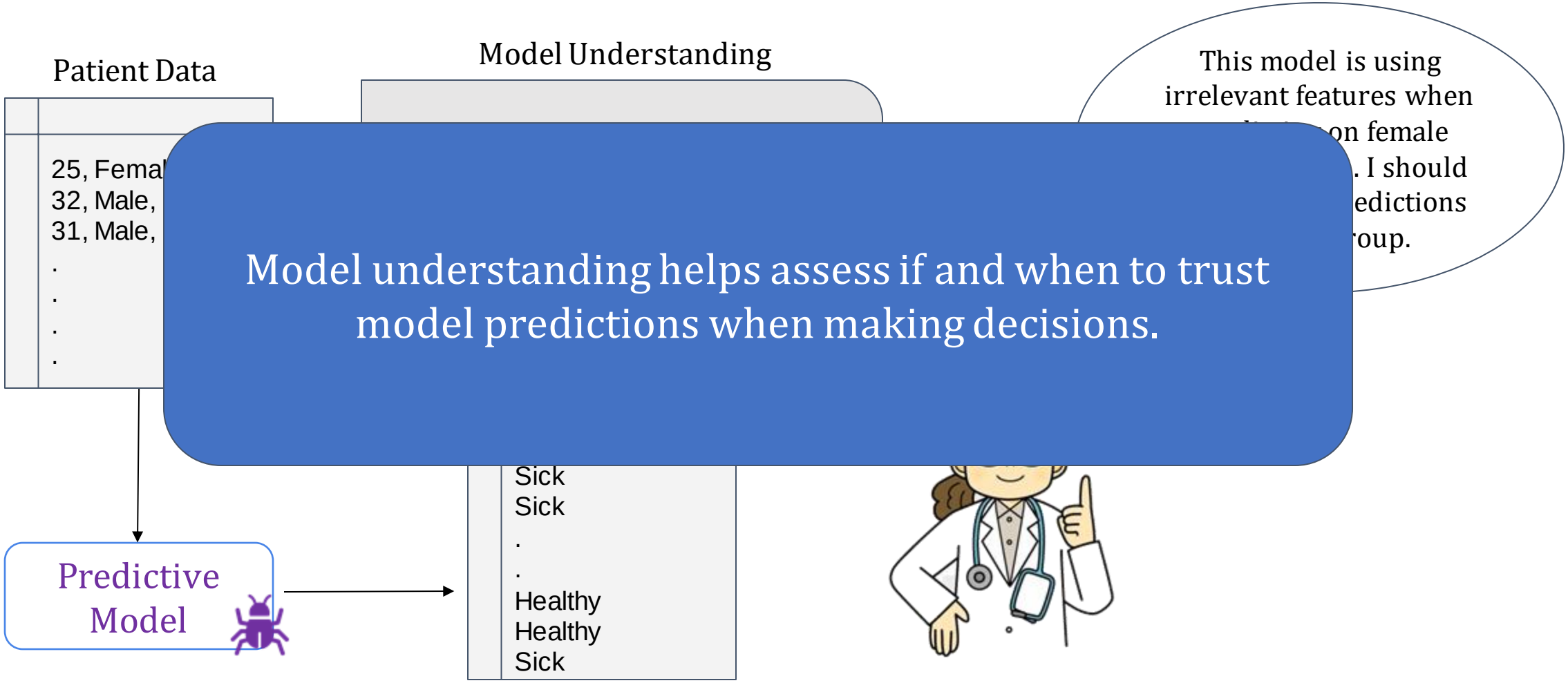
[Larson et. al. 2016]

Motivation: Why Model Understanding?

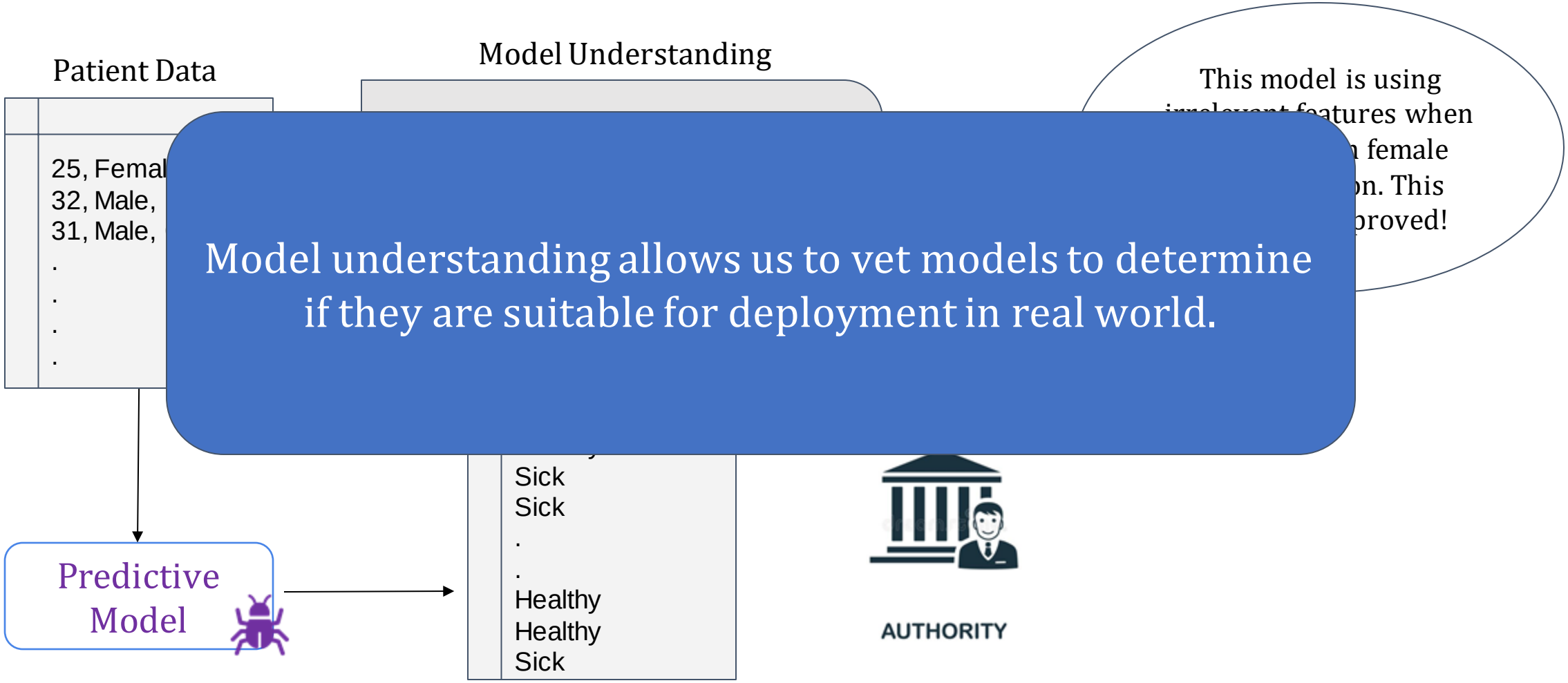
Motivation: Why Model Understanding?



Motivation: Why Model Understanding?



Motivation: Why Model Understanding?



Motivation: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

Decision makers (e.g., doctors, judges)

Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

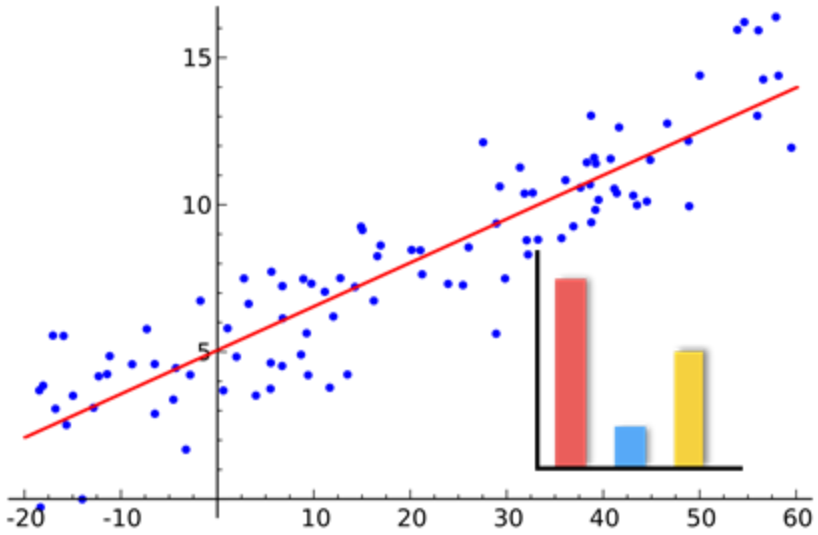
Tutorial Outline

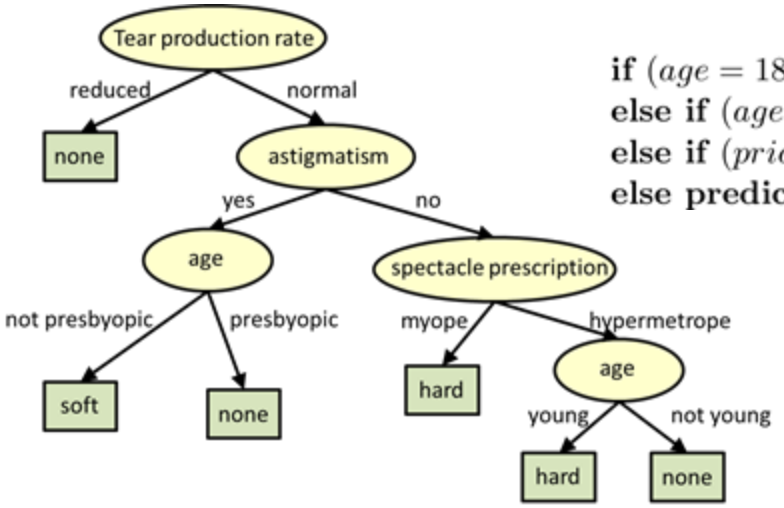
- Motivation
- **Interpretability vs. Explainability**
- Overview of Explanation Methods
- Limitations of Explanation Methods
- Towards Robust & Reliable Explanations
- The Road Ahead

[Letham and Rudin 2015; Lakkaraju et. al. 2016]

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models



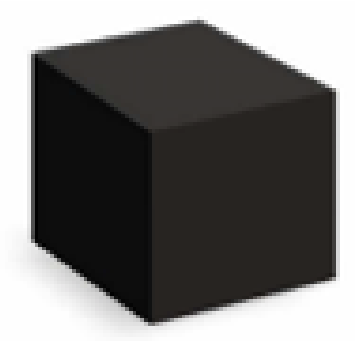


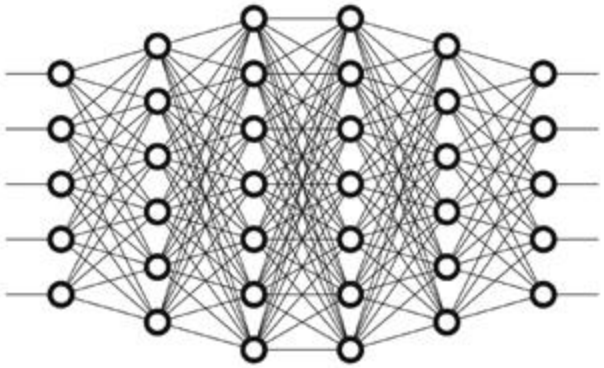
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

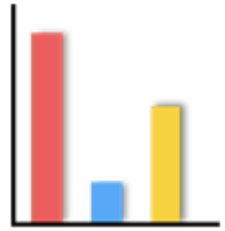
Achieving Model Understanding

Take 2: *Explain* pre-built models *in a post-hoc manner*

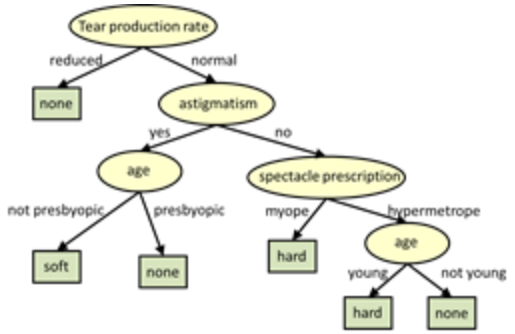




Explainer



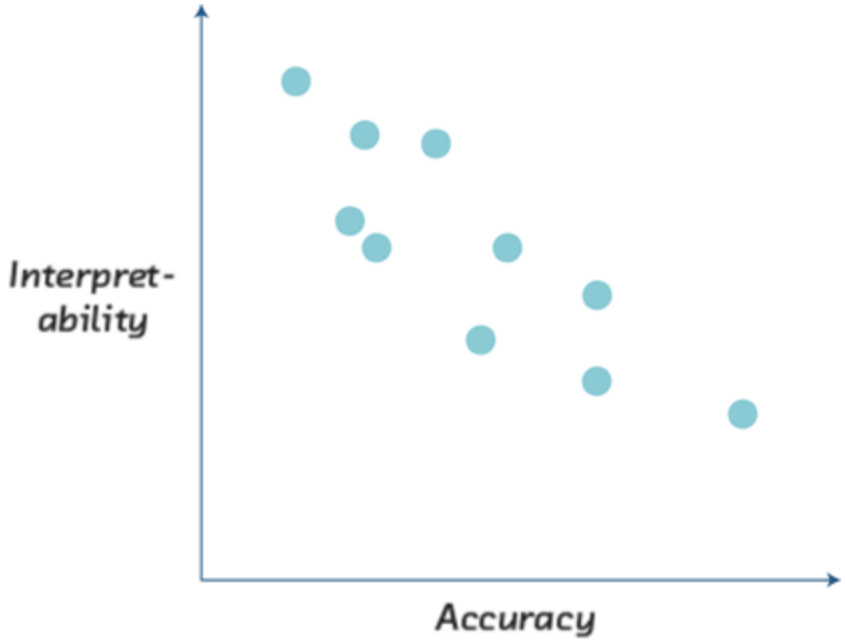
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

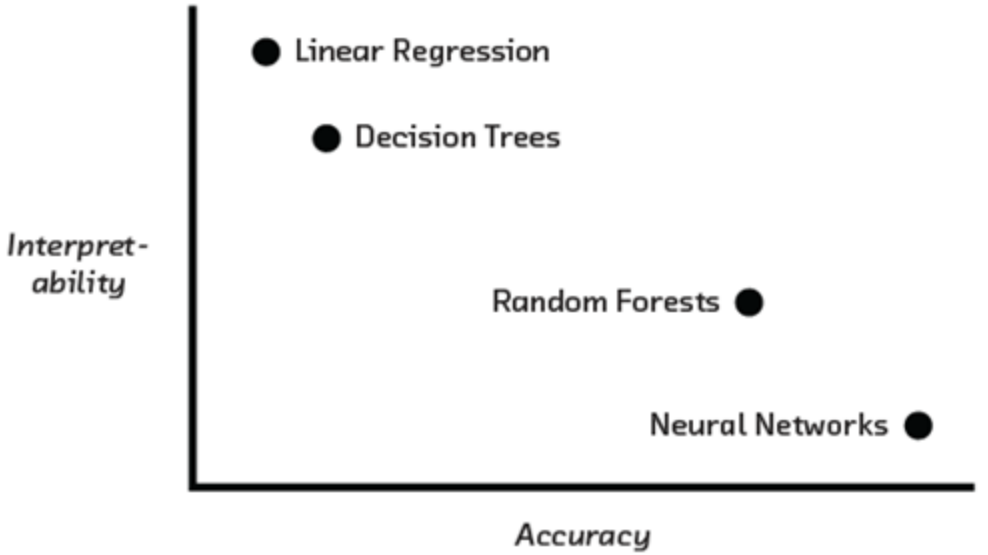


[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations

Example





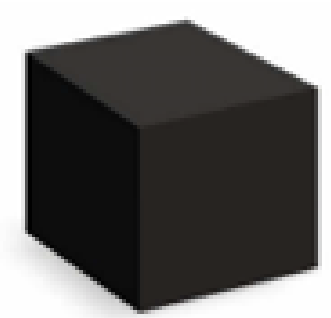
In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

[Ribeiro et. al. 2016]

Inherently Interpretable Models vs. Post hoc Explanations

Sometimes, you don't have enough data to build your model from scratch.

And, all you have is a (proprietary) black box!



[Rudin 2019]

Inherently Interpretable Models vs. Post hoc Explanations

If you *can build* an interpretable model which is also adequately accurate for your setting, DO IT!

Otherwise, ***post hoc explanations*** come to the rescue!

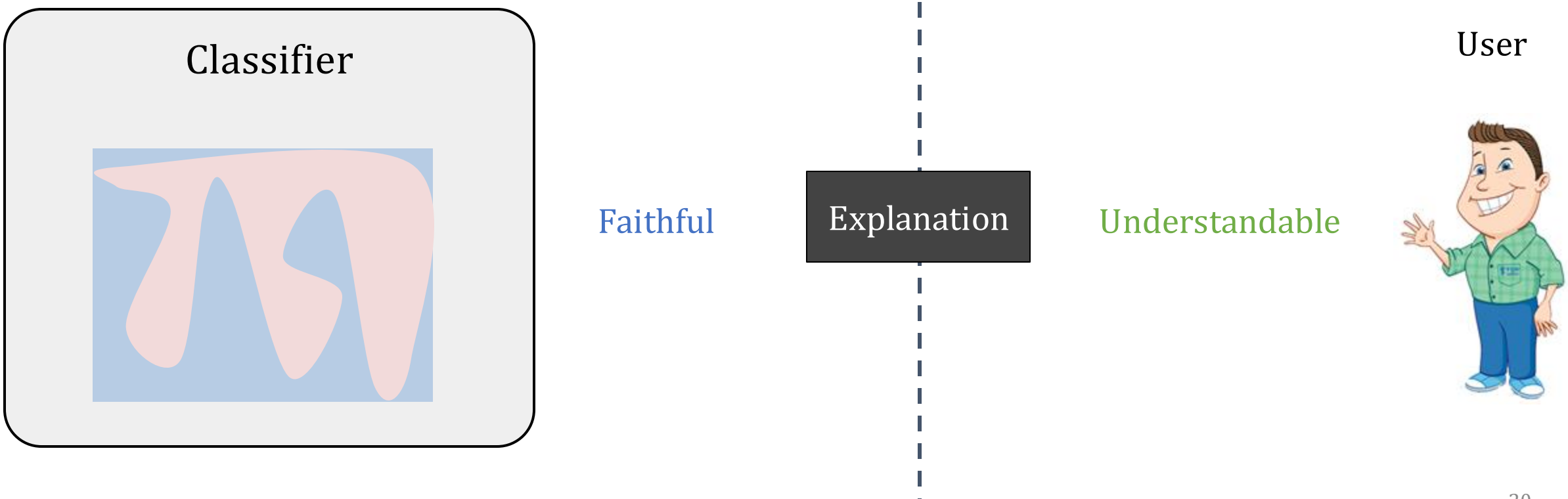
This talk will focus on post hoc explanations!

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- **Overview of Explanation Methods**
- Limitations of Explanation Methods
- Towards Robust & Reliable Explanations
- The Road Ahead

What is an Explanation?

Definition: Interpretable description of the model behavior



Overview of Explanation Methods

Local Explanations vs. Global Explanations

Local Explanations	Global Explanations
Explain individual predictions	Explain complete behavior of the model
Help unearth biases in the *local neighborhood* of a given instance	Sheds light on *big picture biases* affecting larger subgroups
Help vet if individual predictions are being made for the right reasons	Help vet if the model, at a high level, is suitable for deployment

Overview of Explanation Methods

Local Explanations

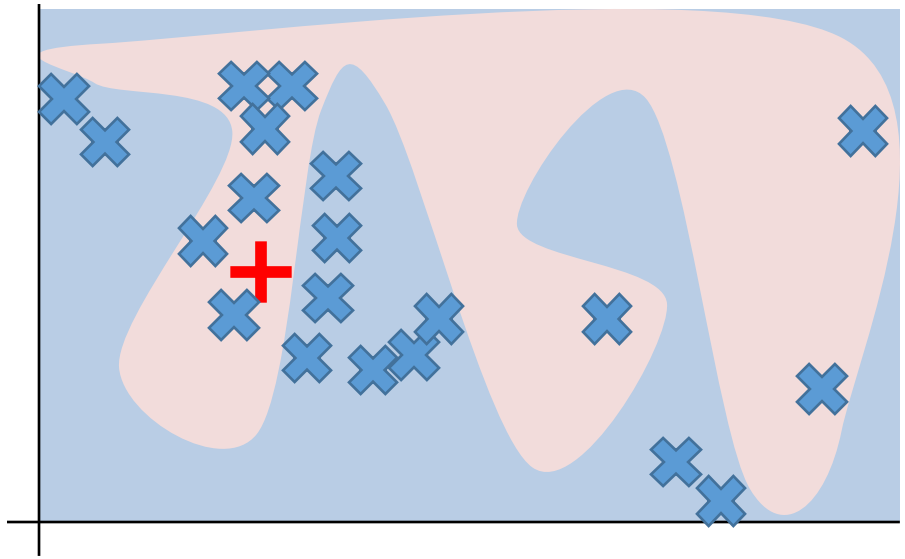
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

LIME: Local Interpretable Model-Agnostic Explanations

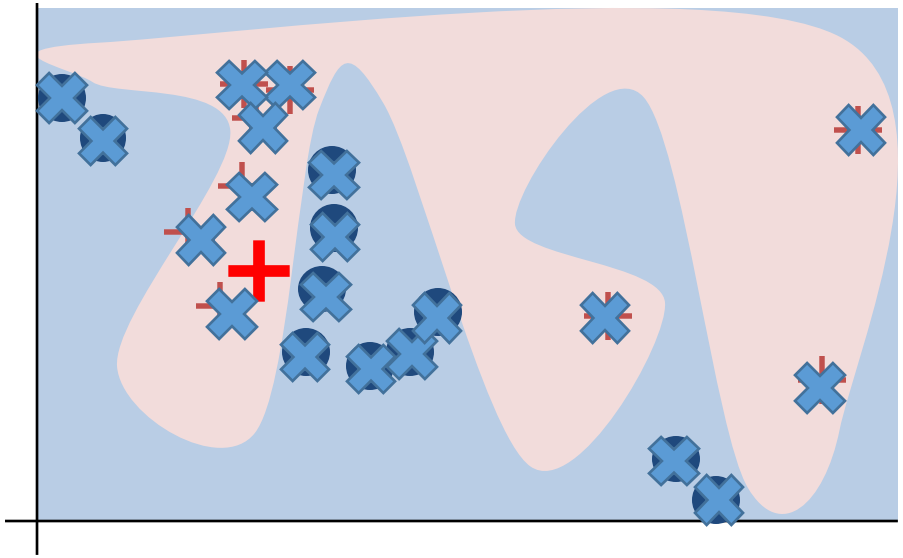
1. Sample points around x_i



[Ribeiro et al. 2016]

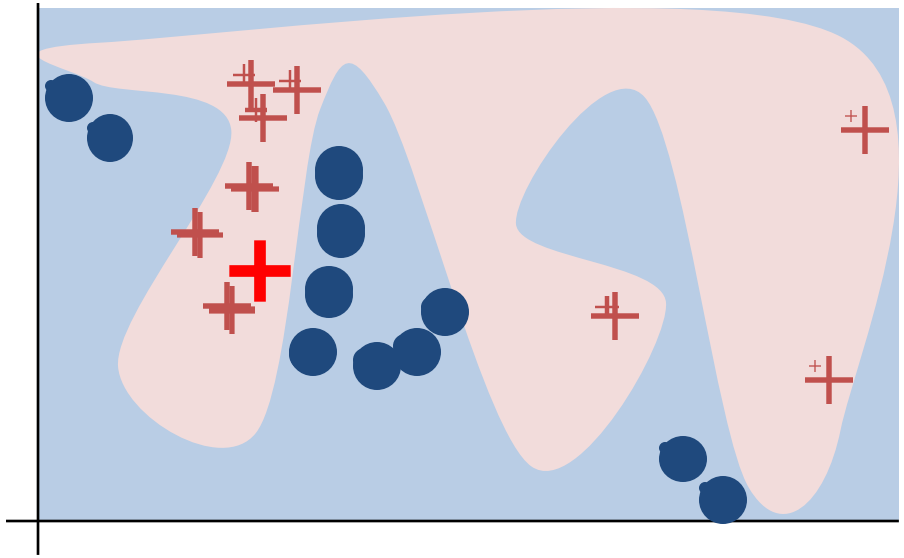
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample



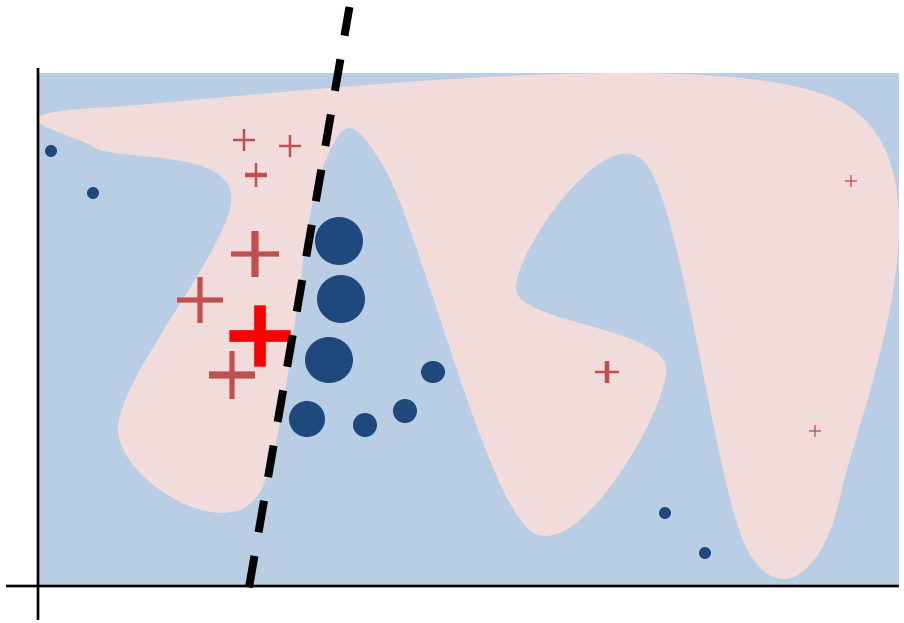
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i



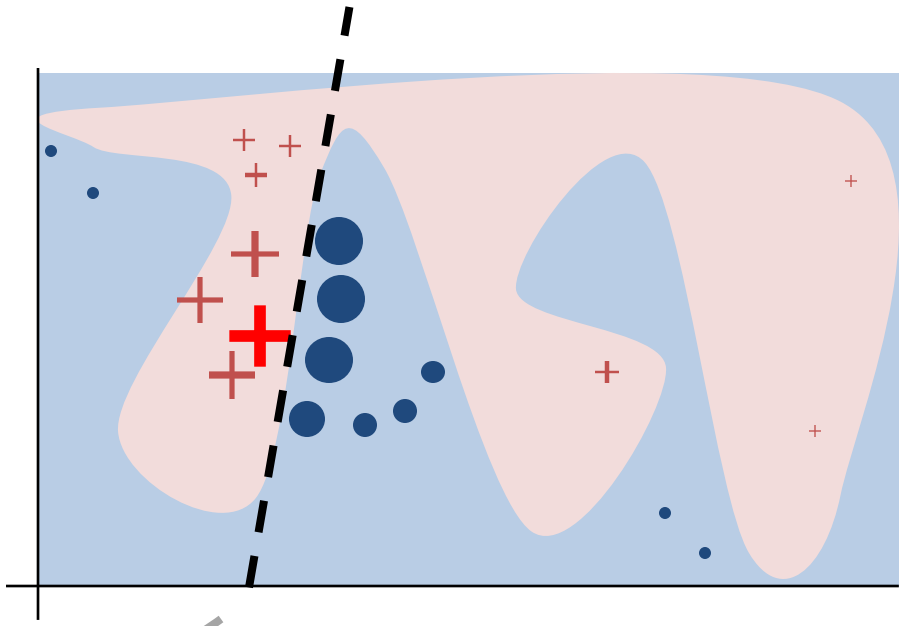
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples



LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples
5. Use simple linear model to explain



Another popular method which outputs feature importances: SHAP

Overview of Explanation Methods

Local Explanations

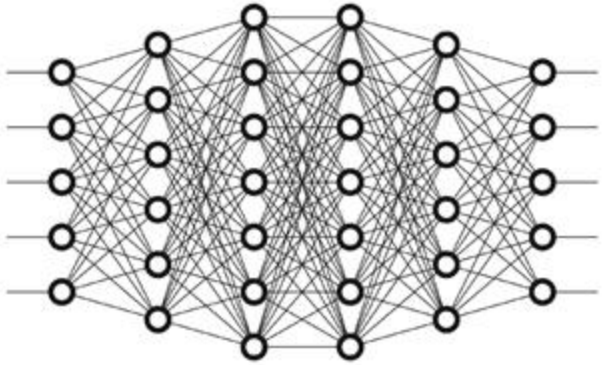
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Saliency Maps

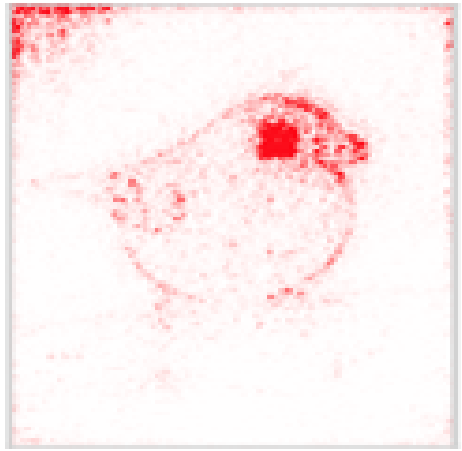
Input



Prediction

Junco Bird

What parts of the input are most relevant for the model's prediction: 'Junco Bird'?



Saliency Map

Saliency Maps

Gradient

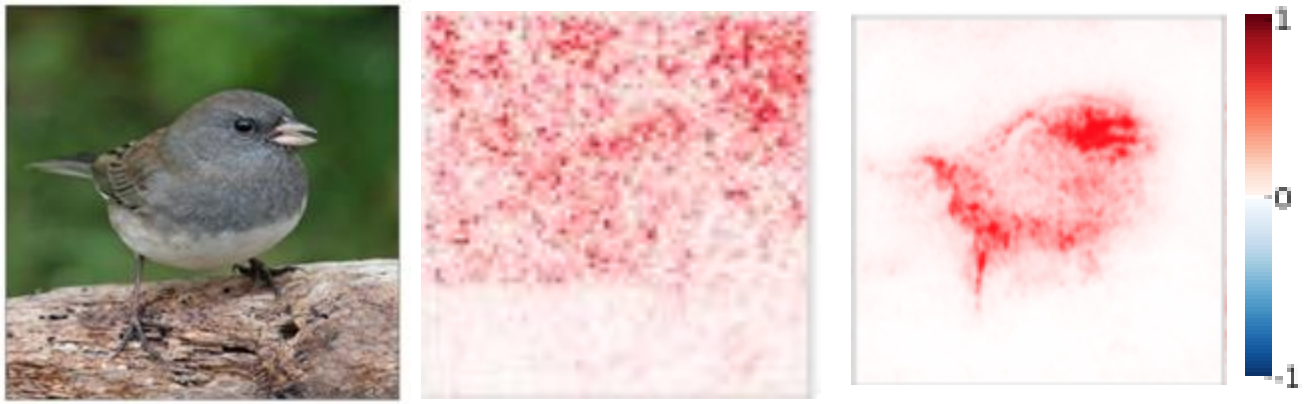
$$\nabla_{\boldsymbol{x}} f$$

Problems:

- noisy and uninterpretable

Saliency Maps

Gradient **SmoothGrad**



$$\frac{1}{n}\sum_{1}^{n}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}+\mathcal{N}(0,\sigma^2))$$

Problems:

- ~~noisy and uninterpretable~~

Overview of Explanation Methods

Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Prototypes/Example

Use examples (synthetic or natural) to explain individual predictions

- Influence Functions (**Koh & Liang 2017**)
  - Identify instances in the training set that are responsible for the prediction of a given test instance

- Activation Maximization (**Erhan et al. 2009**)
  - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest

Overview of Explanation Methods

Local Explanations

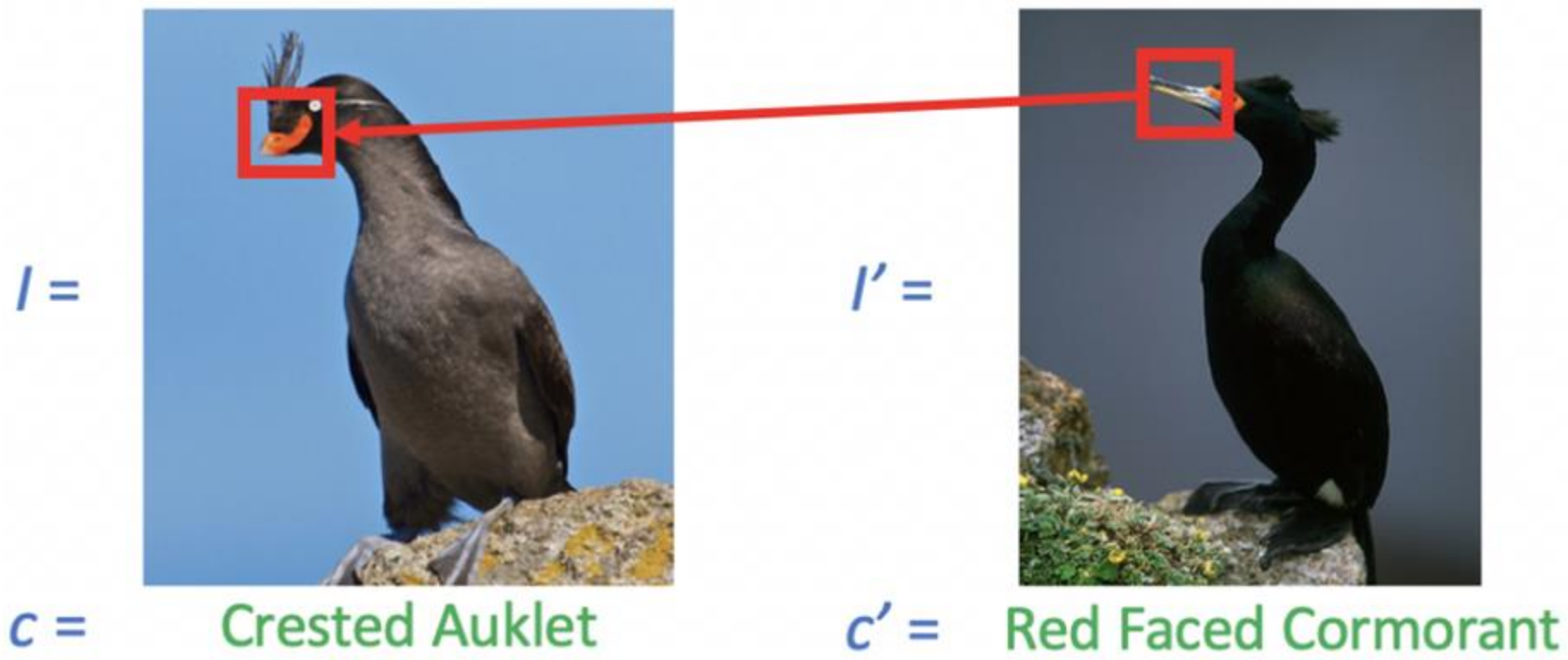
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

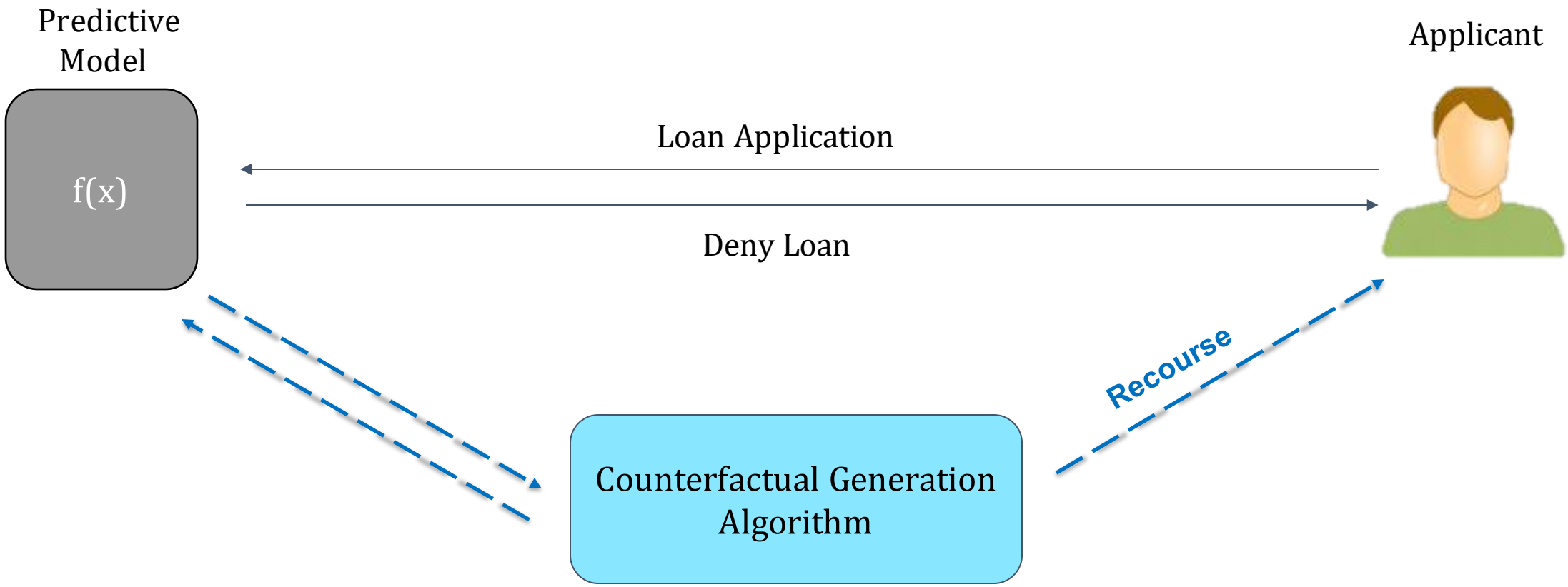
Counterfactual Explanations

What features need to be changed and by how much to flip a model's prediction?



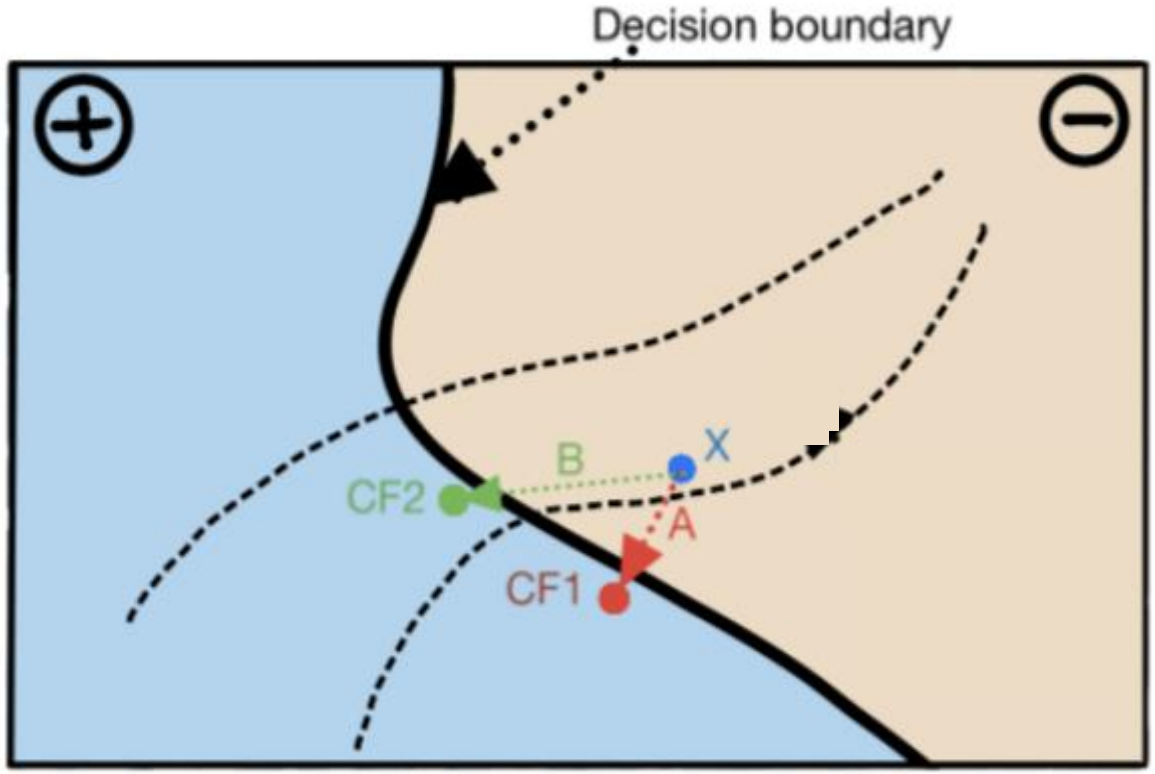
[Goyal et. al., 2019]

Counterfactual Explanations



Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

Generating Counterfactual Explanations: Intuition



Proposed solutions differ on:

1. How to choose among candidate counterfactuals?

1. How much access is needed to the underlying predictive model?

[Verma et. al., 2020]

Overview of Explanation Methods

Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Global Explanations from Local Feature Importances: SP-LIME

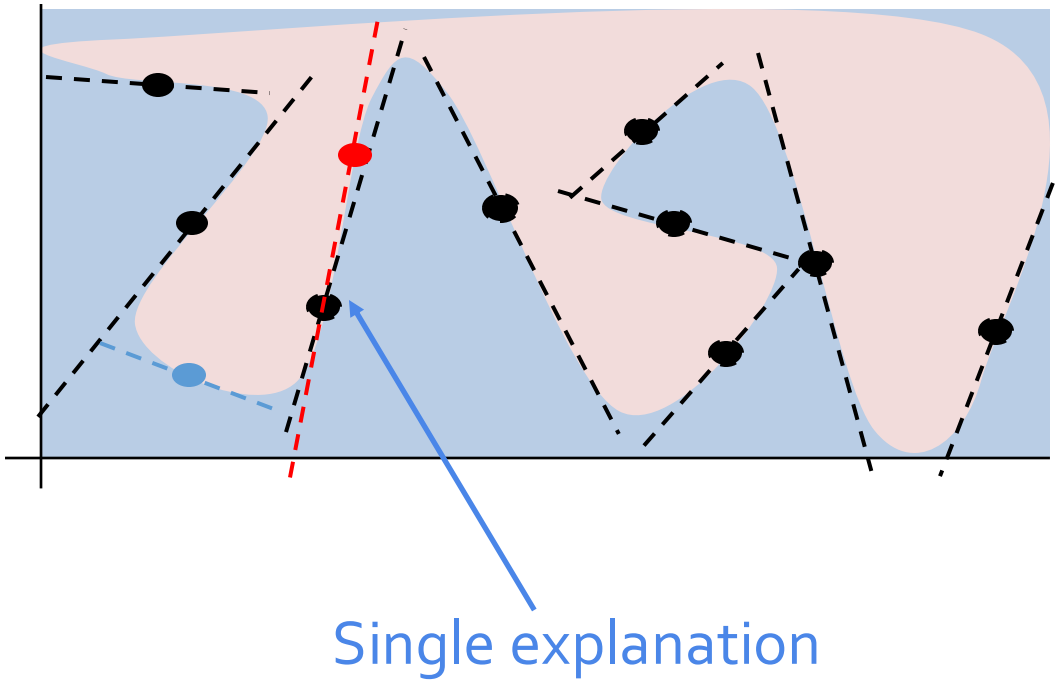
LIME explains a single prediction
local behavior for a single instance

Can't examine all explanations
Instead pick *k* explanations to show to the user

Representative
Should summarize the model's global behavior

Diverse
Should not be redundant in their descriptions

SP-LIME uses submodular optimization and *greedily* picks k explanations



[Ribeiro et. al., 2016]

Overview of Explanation Methods

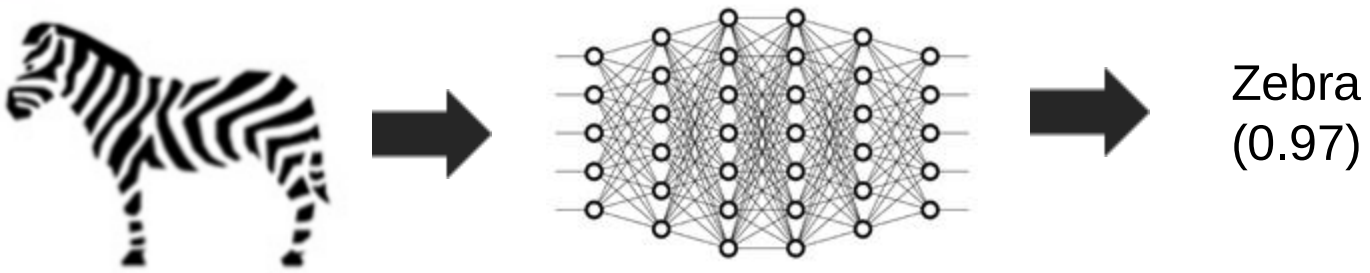
Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Representation Based Explanations



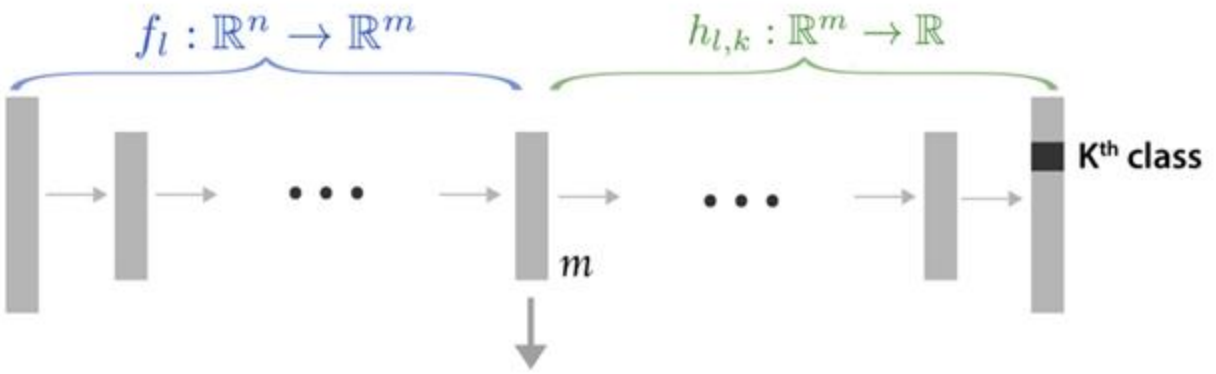
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]

Representation Based Explanations: TCAV

Examples of the concept “stripes”

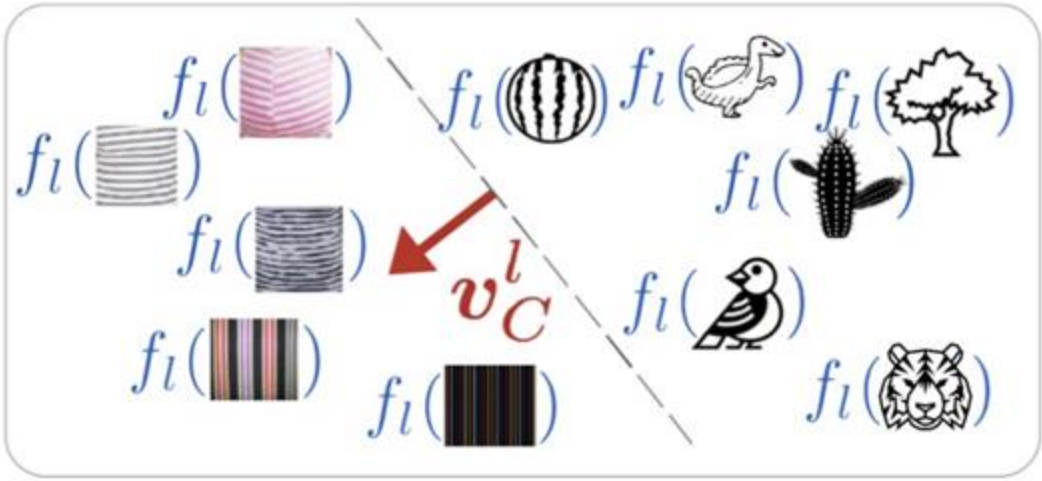
Random examples



Train a linear classifier to separate activations

The vector orthogonal to the decision boundary denotes the concept “stripes”

Compute gradient w.r.t. this vector to determine how important is the notion of stripes for a prediction



Overview of Explanation Methods

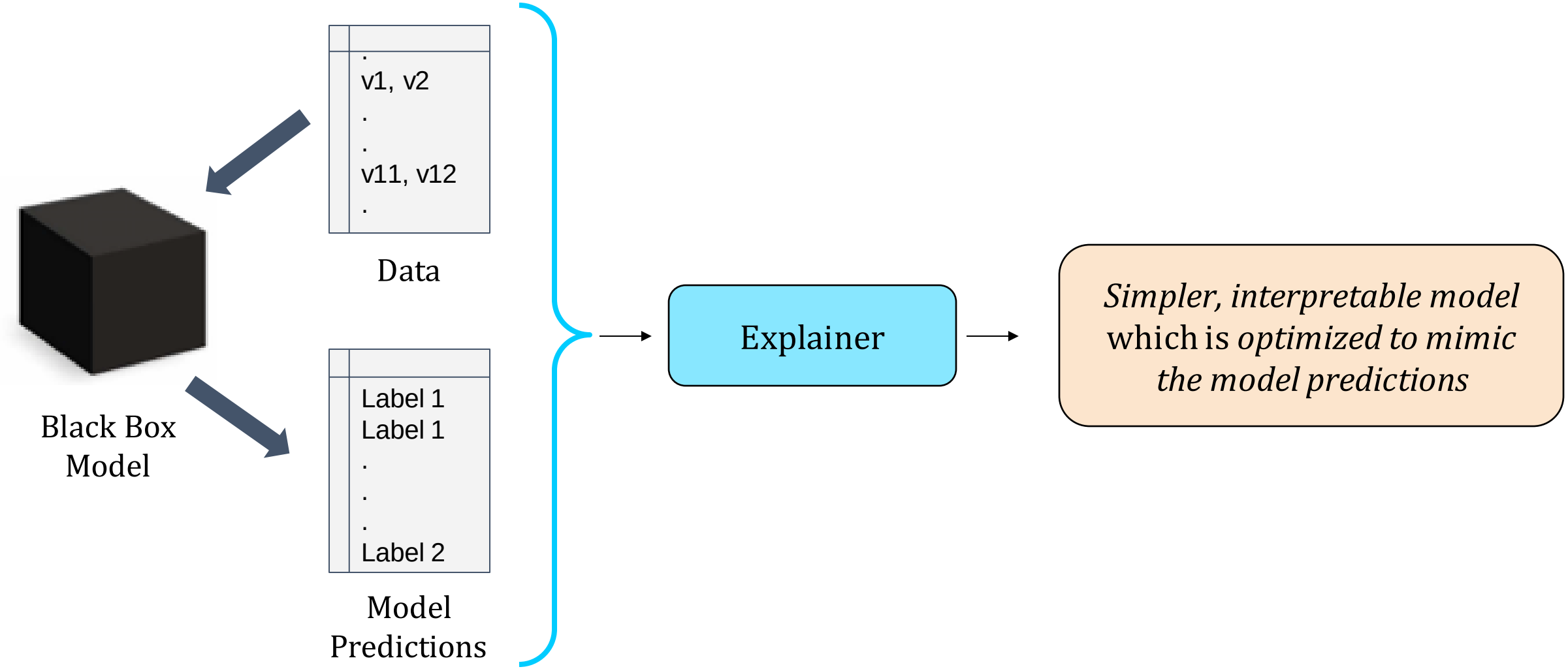
Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

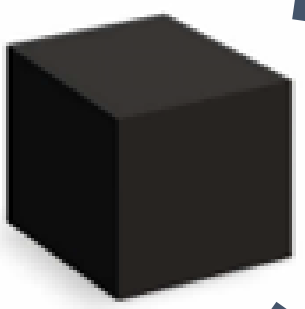
Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Model Distillation



Model Distillation Using Decision Trees



Black Box Model

.
v1, v2
.
.
v11, v12
.

Data

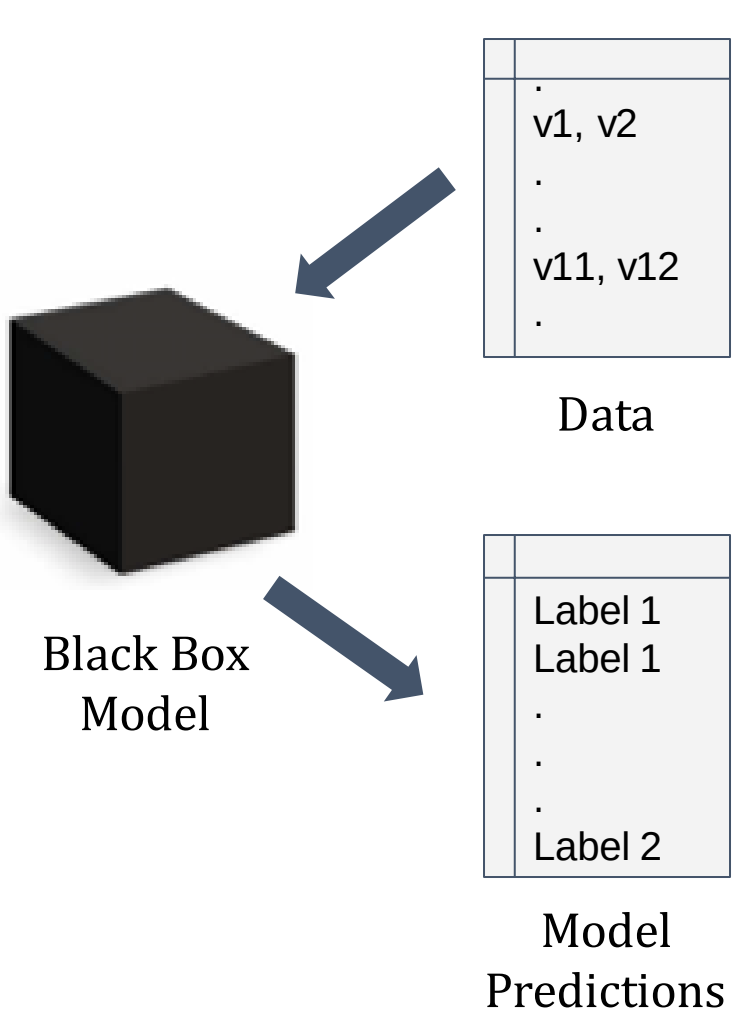
Label 1
Label 1
.
.
.
Label 2

Model Predictions

Explainer

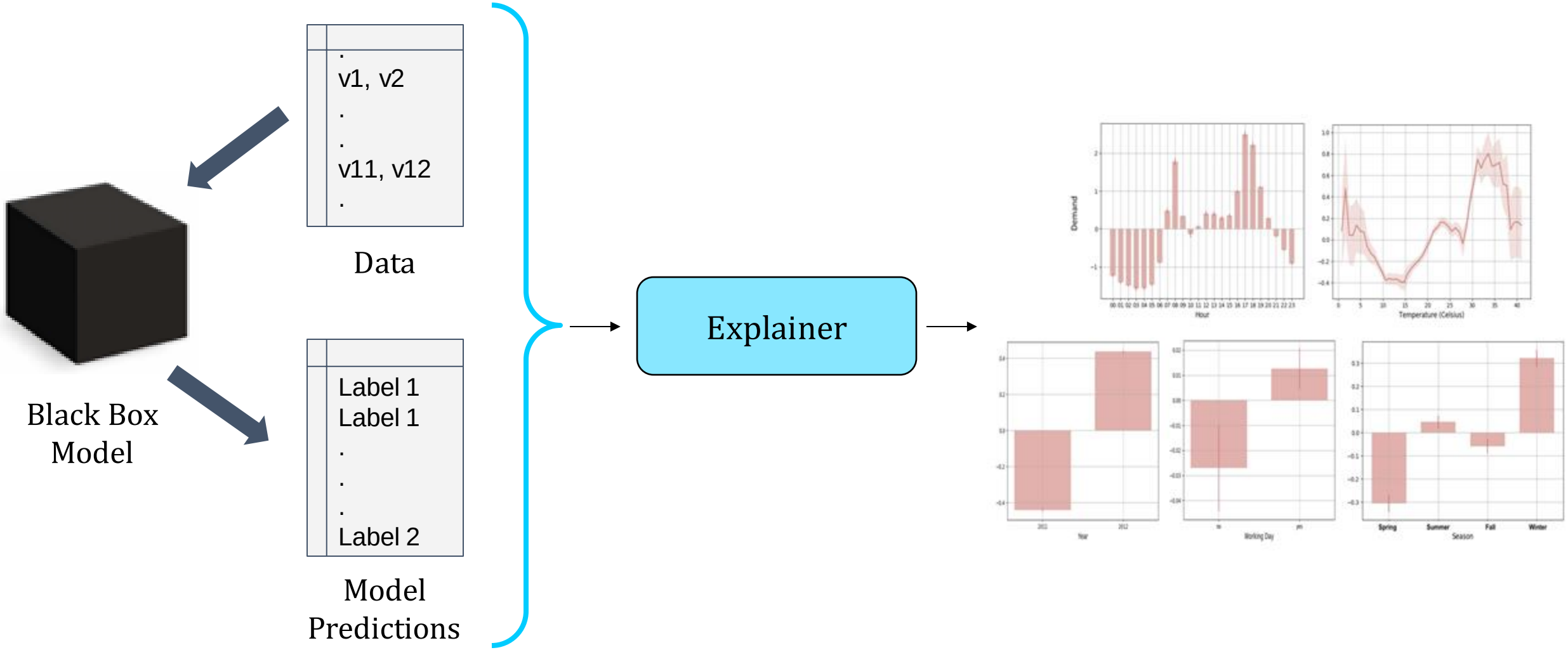
[Bastani et. al., 2019]

Model Distillation Using Decision Sets



[Lakkaraju et. al., 2019]

Model Distillation Using Generalized Additive Models



[Tan et. al., 2019]

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- **Limitations of Explanation Methods**
- Towards Robust & Reliable Explanations
- The Road Ahead

Limitations of Explanation Methods

Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

Fragility
Post-hoc explanations can be easily manipulated.

Limitations of Explanation Methods

Faithfulness

Some explanation methods do not ‘reflect’ the underlying model.

Stability

Slight changes to inputs can cause large changes in explanations.

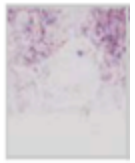
Fragility

Post-hoc explanations can be easily manipulated.

Limitations: **Faithfulness**

Model parameter randomization test

Gradient ⊙ Input



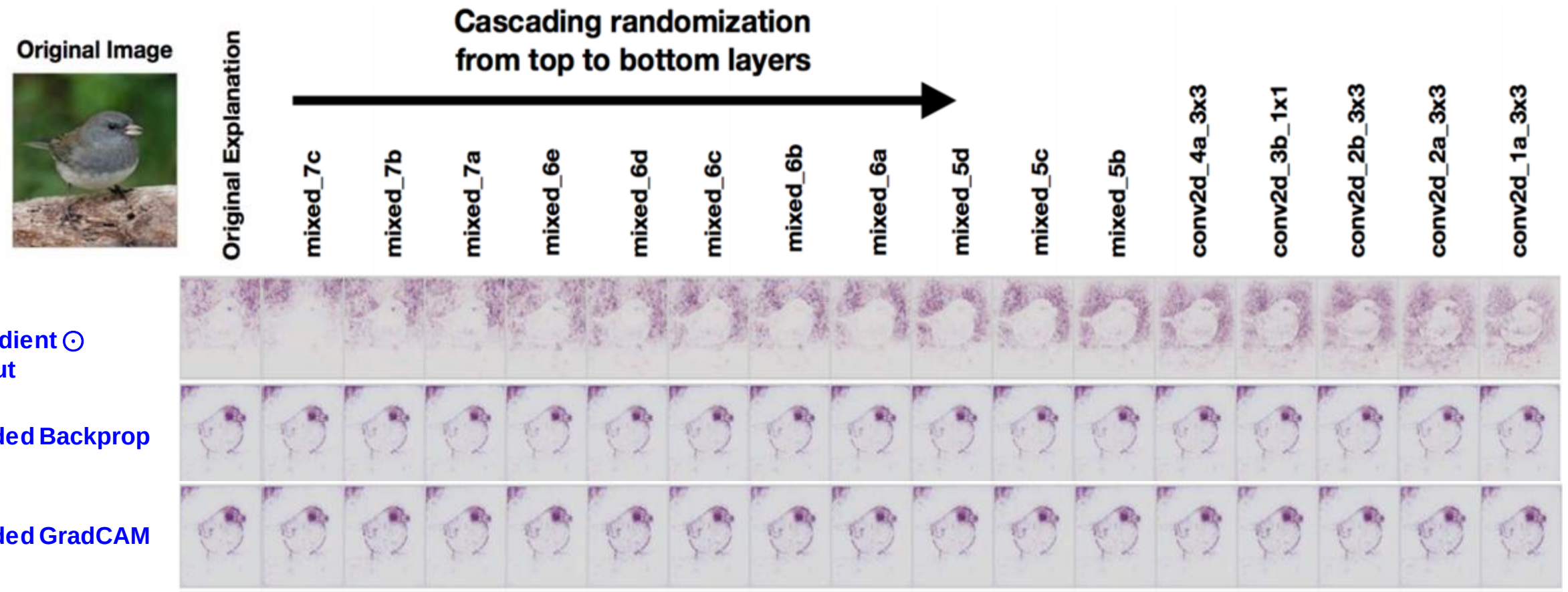
Guided Backprop

Guided GradCAM

Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations: **Faithfulness**

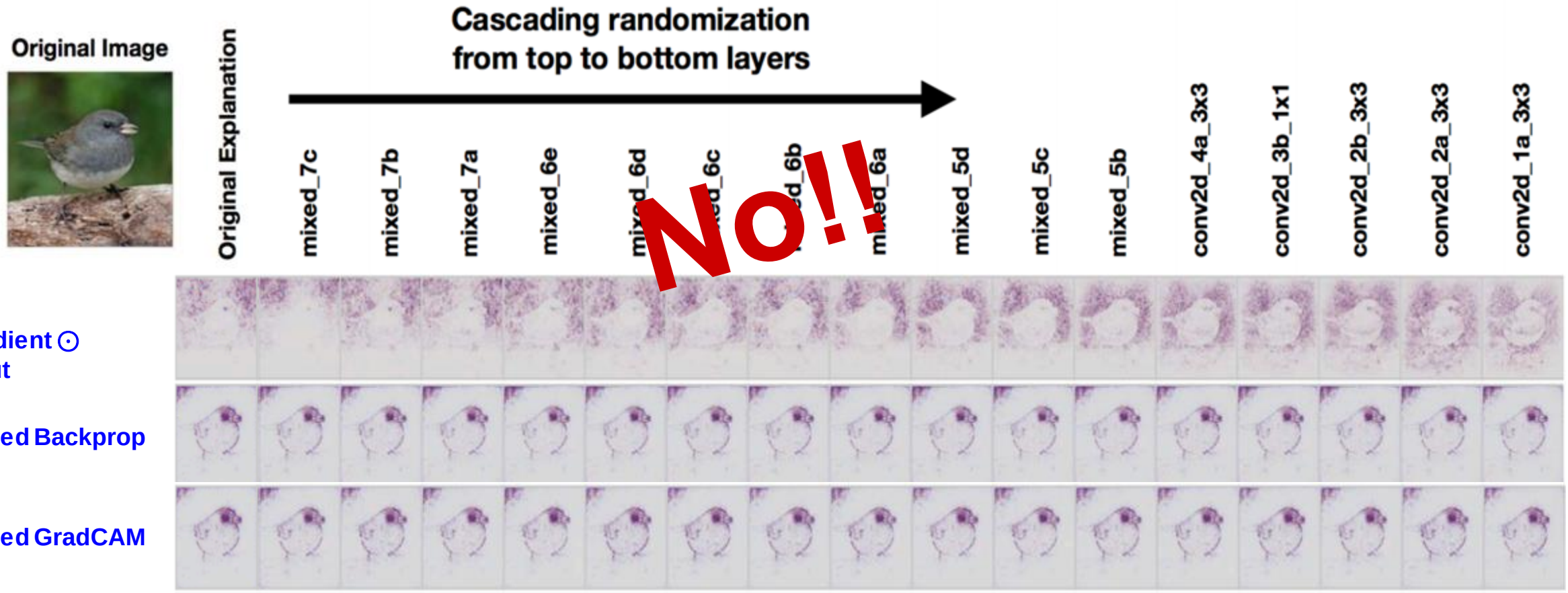
Model parameter randomization test



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations: **Faithfulness**

Model parameter randomization test



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations: **Faithfulness**

Randomizing class labels of instances also didn't impact explanations!

Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations of Explanation Methods

Faithfulness

Some explanation methods do not ‘reflect’ the underlying model.

Stability

Slight changes to inputs can cause large changes in explanations.

Fragility

Post-hoc explanations can be easily manipulated.

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

Local Lipschitz Constant

Explanation function: LIME, SHAP, Gradient...etc.

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

input model

$$A(x, f, H)$$

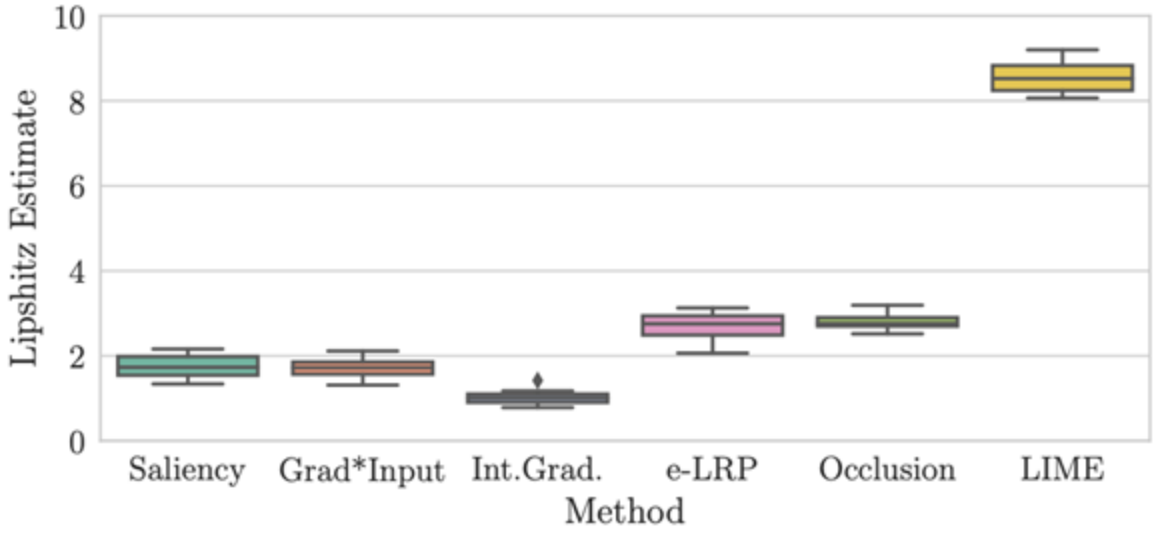
hyperparameters

Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

- Perturbation approaches like LIME can be unstable.

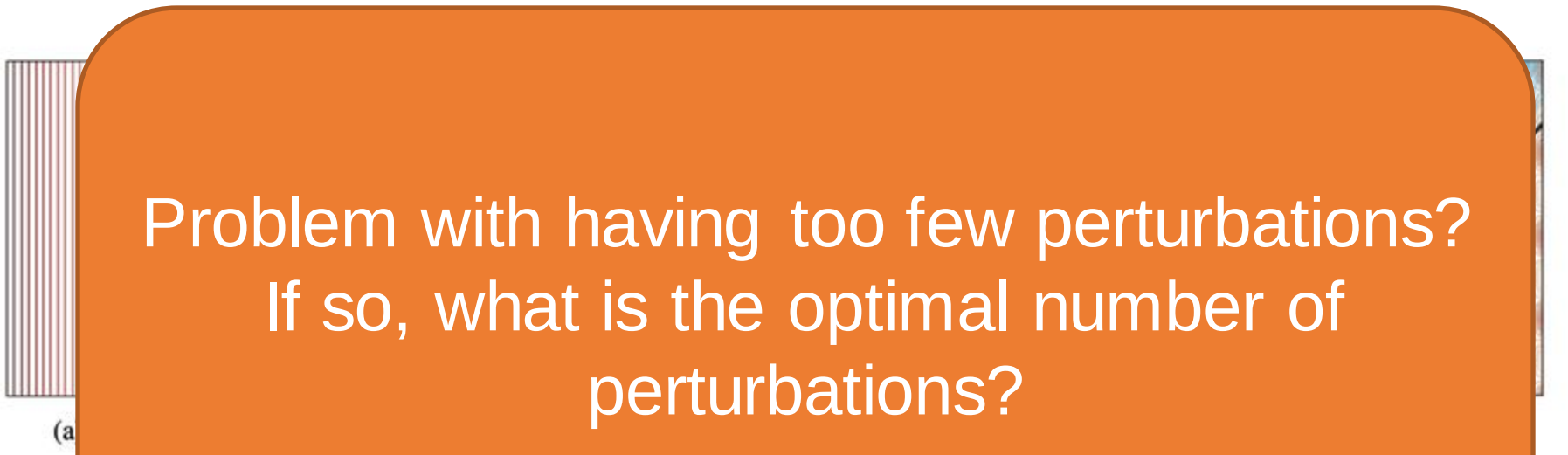


Estimate for 100 tests for an MNIST Model.

Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.

[Slack et. al., 2020]

Limitations: Stability – Problem is Worse!



Problem with having too few perturbations? If so, what is the optimal number of perturbations?

When you repeatedly run LIME on the same instance, you get different explanations (blue region)

Limitations of Explanation Methods

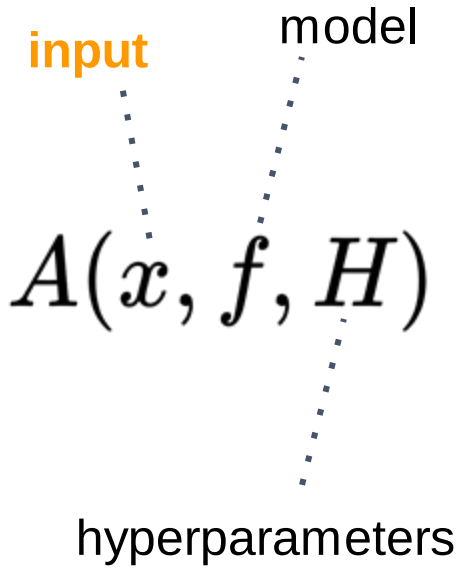
Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

Fragility
Post-hoc explanations can be easily manipulated.

Limitations: **Fragility**

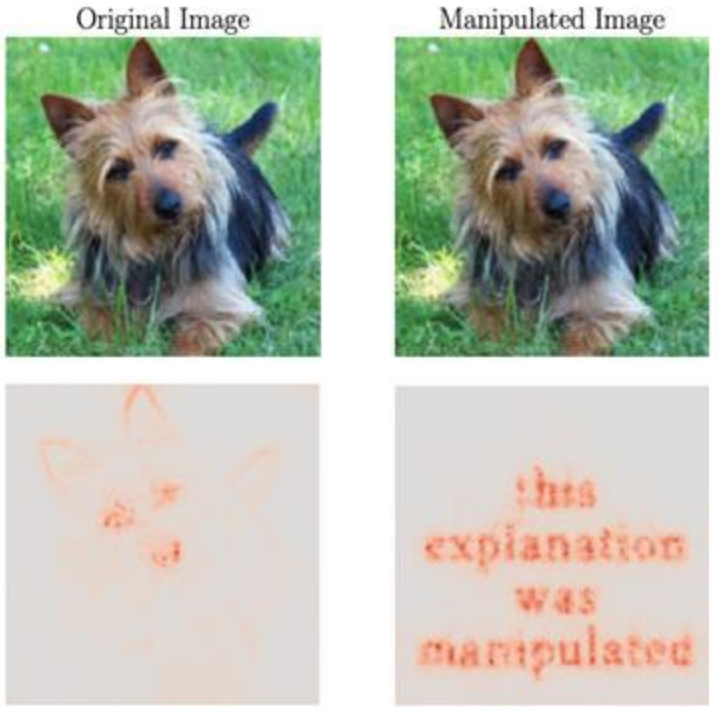
Post-hoc explanations can be easily manipulated

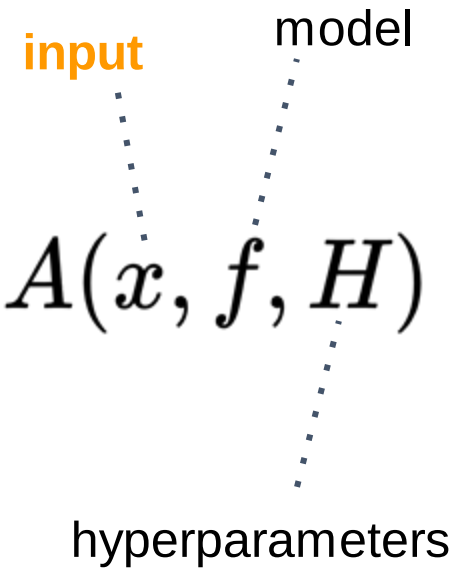


Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." NeurIPS, 2019.

Limitations: **Fragility**

Post-hoc explanations can be easily manipulated

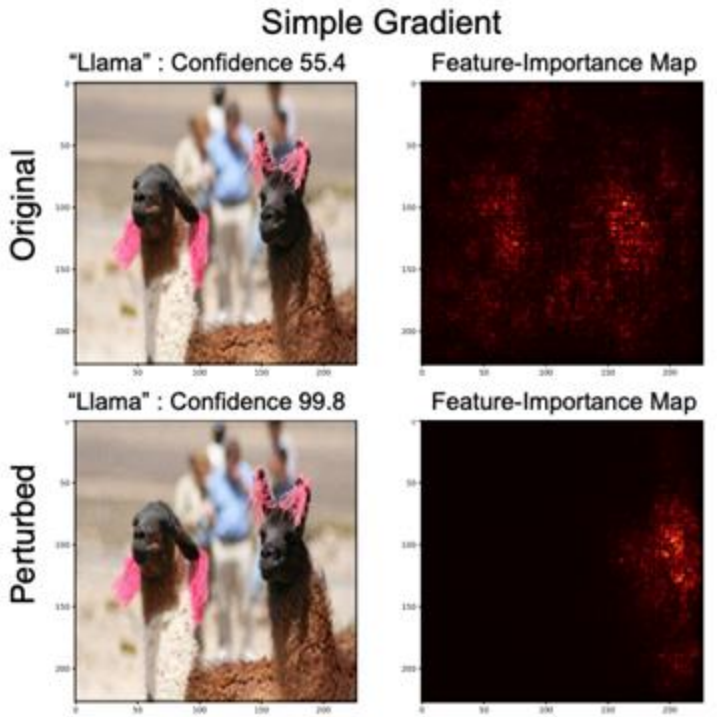




Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." NeurIPS, 2019.

Limitations: **Fragility**

Adversarial Attack on Explanations



$$\arg\max_{\delta} \mathcal{D}(\mathbf{I}(\mathbf{x}_t; \mathcal{N}), \mathbf{I}(\mathbf{x}_t + \delta; \mathcal{N}))$$

$$\text{subject to: } ||\delta||_\infty \leq \epsilon$$

$$f(\mathbf{x}_t + \delta) = f(\mathbf{x})$$

adversarial input

model

$$A(x, f, H)$$

hyperparameters

Ghorbani, Amirata et al. "Interpretation of neural networks is fragile." AAAI, 2019.

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- **Towards Robust & Reliable Explanations**
- The Road Ahead

[Lakkaraju et. al., 2020]

RObust & Stable Post hoc Explanations (ROPE)

- Framework for generating explanations that are stable and robust to distribution shifts

- It is flexible, e.g., it can be instantiated for linear vs. rule based explanations

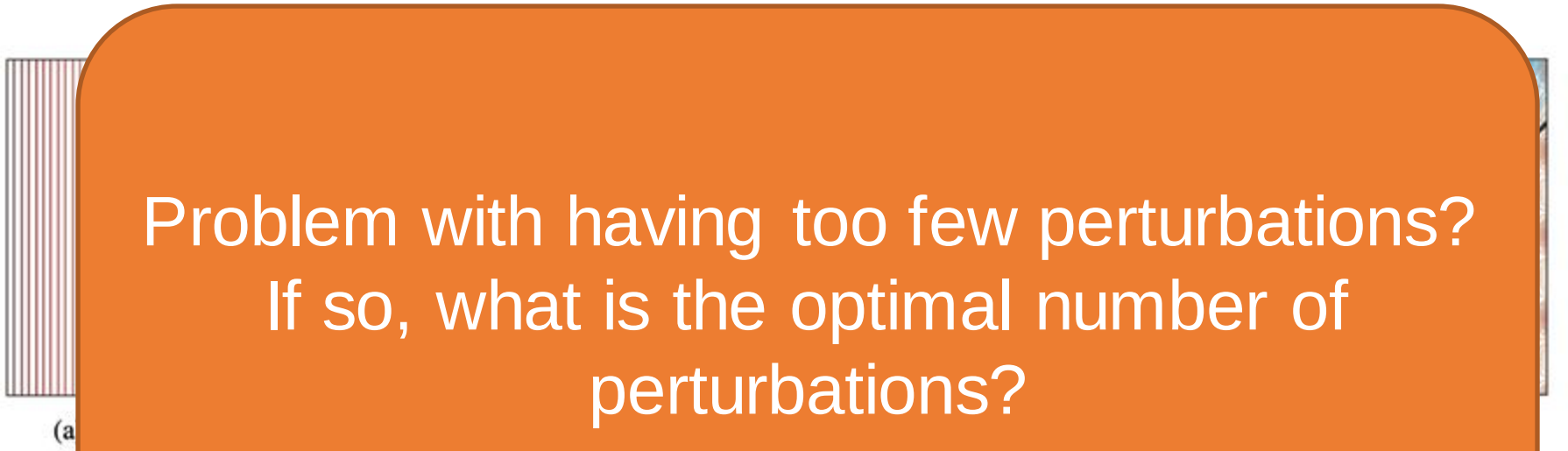
$$\hat{E} = \arg\min_{E \in \mathcal{E}} \max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))].$$

worst-case computed over plausible distribution shifts

expected gap between explanation and black box

[Slack et. al., 2020]

Limitations: Stability – Problem is Worse!

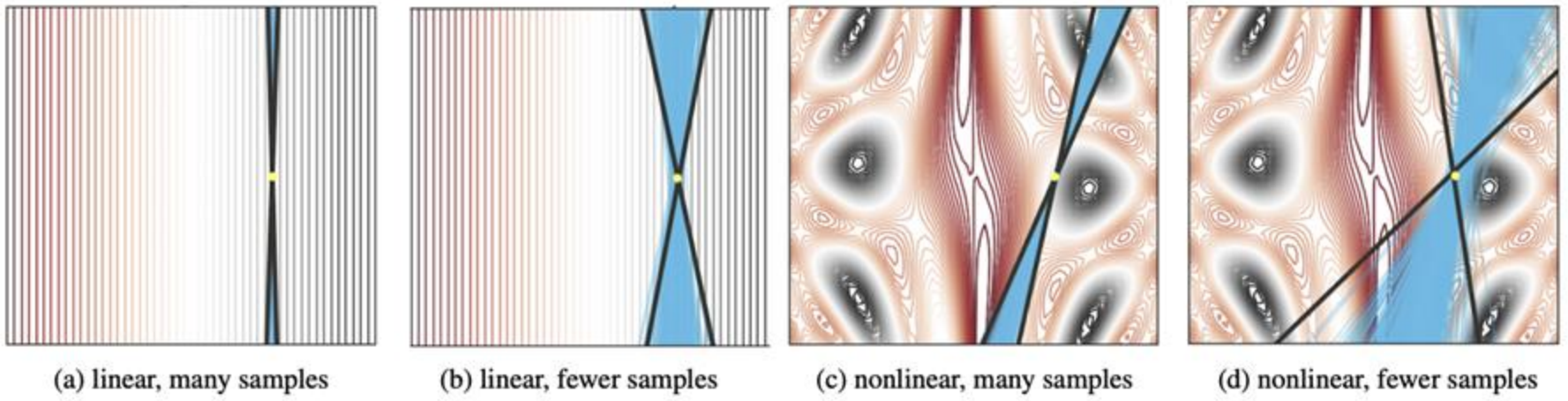


Problem with having too few perturbations? If so, what is the optimal number of perturbations?

When you repeatedly run LIME on the same instance, you get different explanations (blue region)

[Slack et. al., 2020]

Modeling Uncertainty of Black Box Explanations: BayesLIME & BayesSHAP



BayesLIME 95% Confidence Interval Shown by Black Lines

Modeling Uncertainty of Black Box Explanations: BayesLIME & BayesSHAP

$$y|z,\phi,\sigma^2 \sim \phi^T z + \underbrace{\mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})}_{\epsilon}, \quad \forall z \in \mathcal{Z}$$

$$\phi|\sigma^2 \sim \mathcal{N}(\phi_0, \sigma^2\Sigma_0) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2).$$

Proximity function

Feature importances

- No need to resort to MCMC or VI; Closed form solutions

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- Towards Robust & Reliable Explanations
- **The Road Ahead**

The Road Ahead

- Explainability as a technology is fragile; Research is in progress
- Improving the Reliability of Explanations
- Developing Evaluation Frameworks for Explanations
- Focusing on the Scalability of Explanation Methods

Thank You!

- Email: hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu;
- Course on interpretability and explainability: https://interpretable-ml-class.github.io/
- Trustworthy ML Initiative: https://www.trustworthyml.org/
  - Lots of resources and seminar series on topics related to explainability, fairness, adversarial robustness, differential privacy, causality etc.